Assignment 3. Neural Network Classifier with Two Hidden Layers

In this assignment, three neural network classifiers with two hidden layers for binary classification has been built: one making the prediction based on the first two features, another one predicting based on the first three features and lastly, a classifier based on four features. The data used for this assignment is the banknote authentication data set. There is a total of four features with 1372 examples, each example representing a banknote.

The activation function at the hidden units is ReLU. The network is trained using the average cross-entropy loss on the training set and stochastic gradient descent (SGD) is used in the optimization process. **The learning rate used is 0.001. The number of epochs is fixed at 100. The initialization weights** chosen to achieve the smallest validation misclassification error overall is labelled above each table.

Below show three tables, one for each D ∈ {2,3,4} containing validation error for each pair (n_1,n_2). The tables also include the test misclassification error for the best classifier for the corresponding D.

D = 2

Weights were initialized 0.3 to achieve the smallest validation misclassification error.

(n1, n2)	Validation Misclass. Error	Test Misclass. Error
(2,2)	0.1245136186770428	-
(2,3)	0.12062256809338522	0.07871720116618076
(2,4)	0.12062256809338522	-
(3,2)	0.12062256809338522	-
(3,3)	0.12062256809338522	-
(3,4)	0.12062256809338522	-
(4,2)	0.12062256809338522	-
(4,3)	0.12062256809338522	-
(4,4)	0.12062256809338522	-

The number of units chosen in the hidden layer is (2,3) with weights at:

W1: [0.47957291 0.82457231 0.44092613]
[0.47957291 0.82457231 0.44092613]

W2: [-0.00089109 0.60570547 0.60570547]
[-0.00089109 0.60570547 0.60570547]
[-0.00089109 0.60570547 0.60570547]

W3: [0.96556734 -0.76371108 -0.76371108 -0.76371108]

D = 3

Weights were initialized 0.7 to achieve the smallest validation misclassification error.

(n1, n2)	Validation Misclass. Error	Test Misclass. Error
(2,2)	0.019455252918287938	0.03498542274052478
(2,3)	0.04669260700389105	-
(2,4)	0.07003891050583658	-
(3,2)	0.03501945525291829	-
(3,3)	0.042801556420233464	-
(3,4)	0.07392996108949416	-
(4,2)	0.042801556420233464	-
(4,3)	0.06614785992217899	-
(4,4)	0.0933852140077821	-

The number of units chosen in the hidden layer is (2,2) with weights at:

W1: [0.42490467 0.57380875 0.65766382 0.70404154]
[0.42490467 0.57380875 0.65766382 0.70404154]

W2: [0.45319586 0.48014502 0.48014502]
[0.45319586 0.48014502 0.48014502]

W3: [0.1717457 -0.14571835 -0.14571835]

D = 4.

Weights were initialized 0.7 to achieve the smallest validation misclassification error.

(n1,n2)	Validation Misclass. Error	Test Misclass. Error
(2,2)	0.10116731517509728	-
(2,3)	0.08949416342412451	-
(2,4)	0.07392996108949416	0.09620991253644315
(3,2)	0.10116731517509728	-
(3,3)	0.08949416342412451	-
(3,4)	0.08560311284046693	-
(4,2)	0.10505836575875487	-
(4,3)	0.08560311284046693	-
(4,4)	0.10116731517509728	-

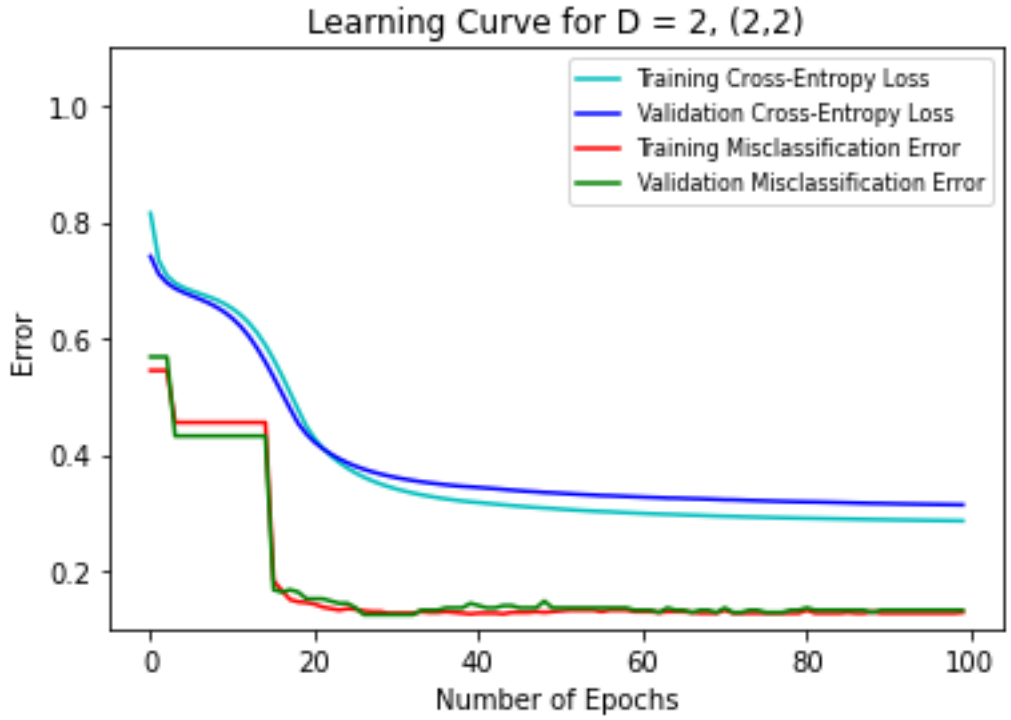
The number of units chosen in the hidden layer is (2,4) with weights at:

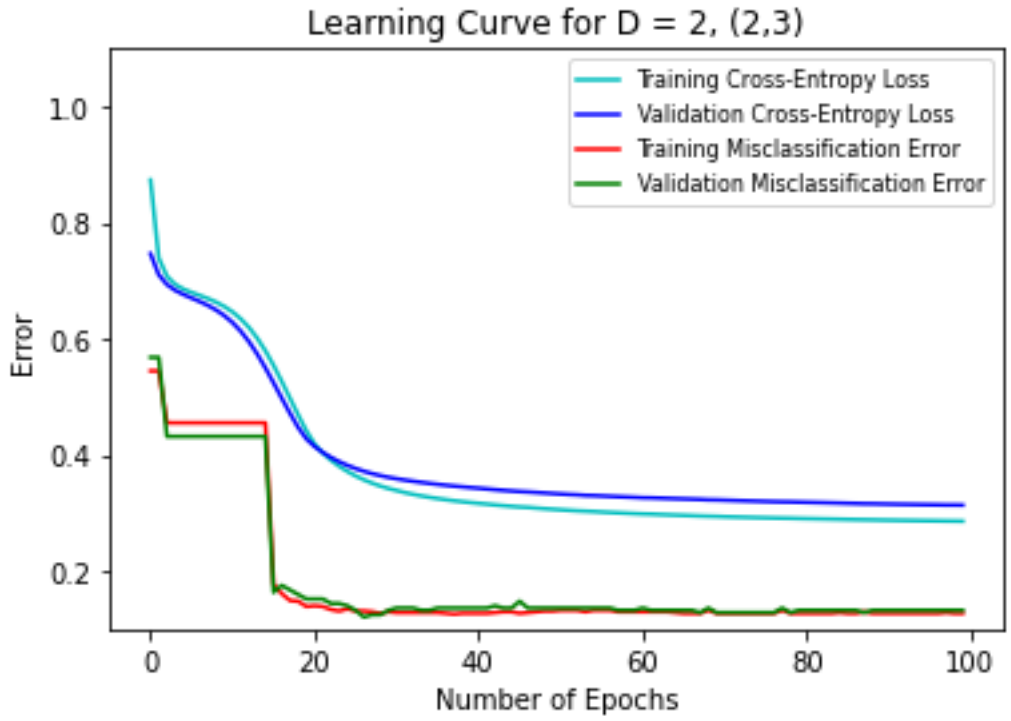
W1: [0.09353236 0.52694249 0.74567599 0.63413137 0.50687609]
[0.09353236 0.52694249 0.74567599 0.63413137 0.50687609]

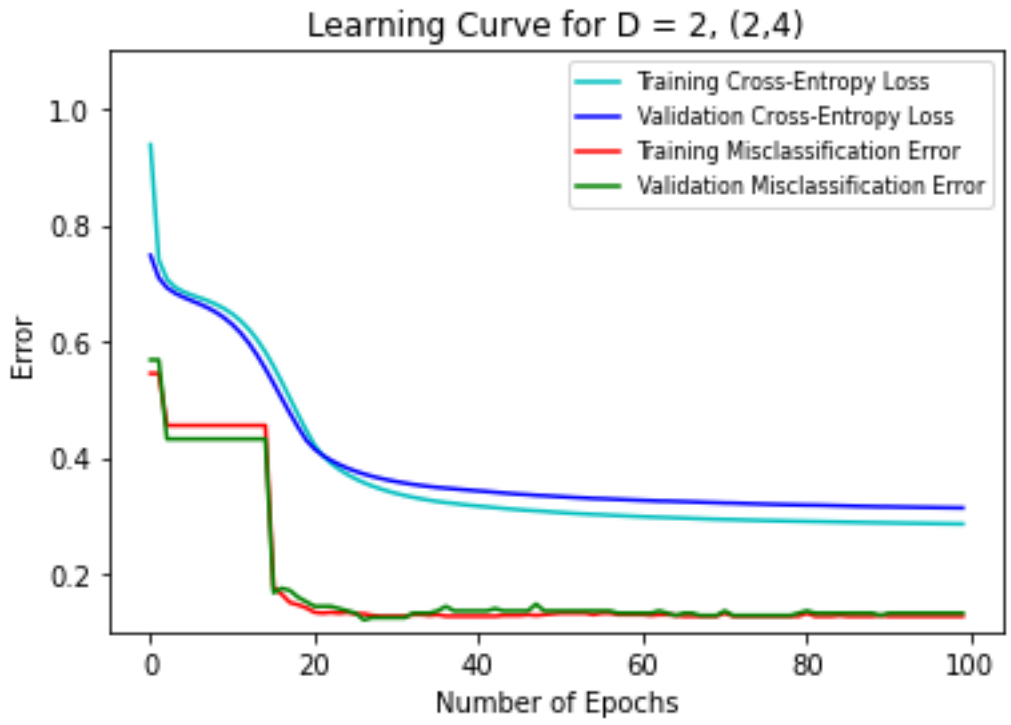
W2: [0.42919854 0.45177311 0.45177311]
[0.42919854 0.45177311 0.45177311]
[0.42919854 0.45177311 0.45177311]

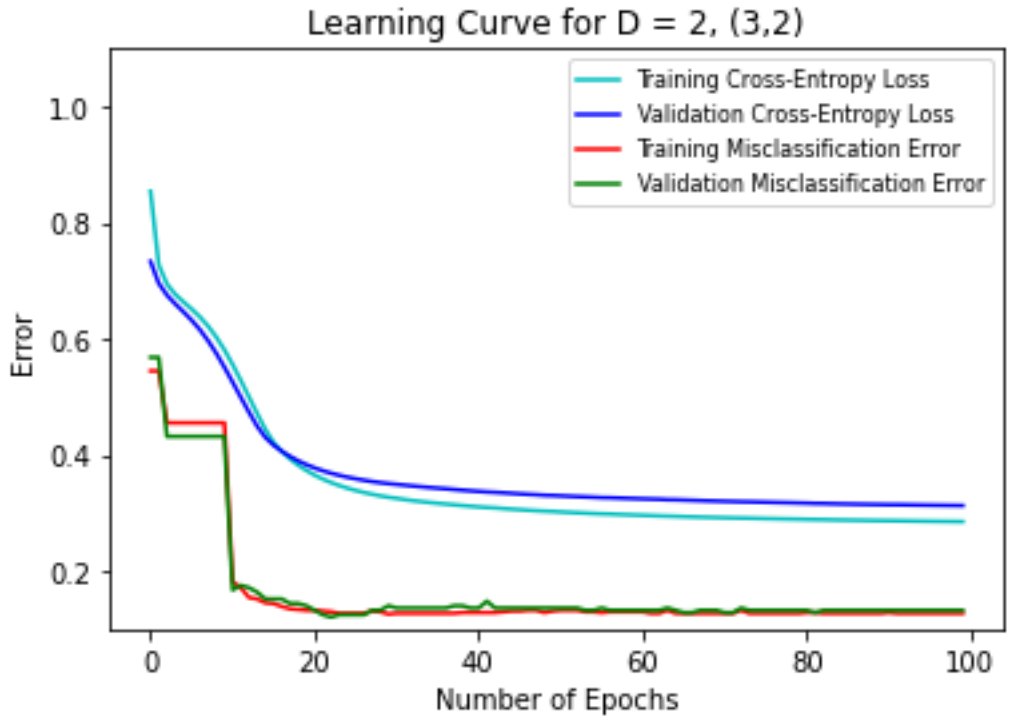
[0.42919854 0.45177311 0.45177311]
W3: [0.14443956 -0.06849543 -0.06849543 -0.06849543 -0.06849543]

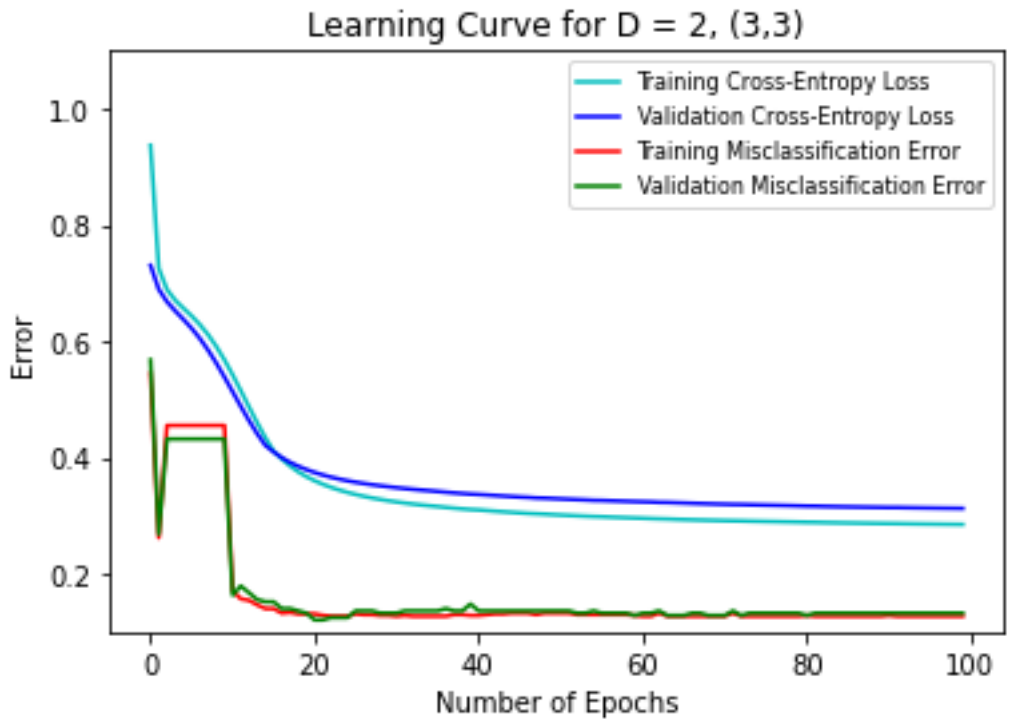
Below show the plots of the training and validation cross-entropy losses and misclassification errors versus the epoch number, obtained while running the SGD for D = 2.

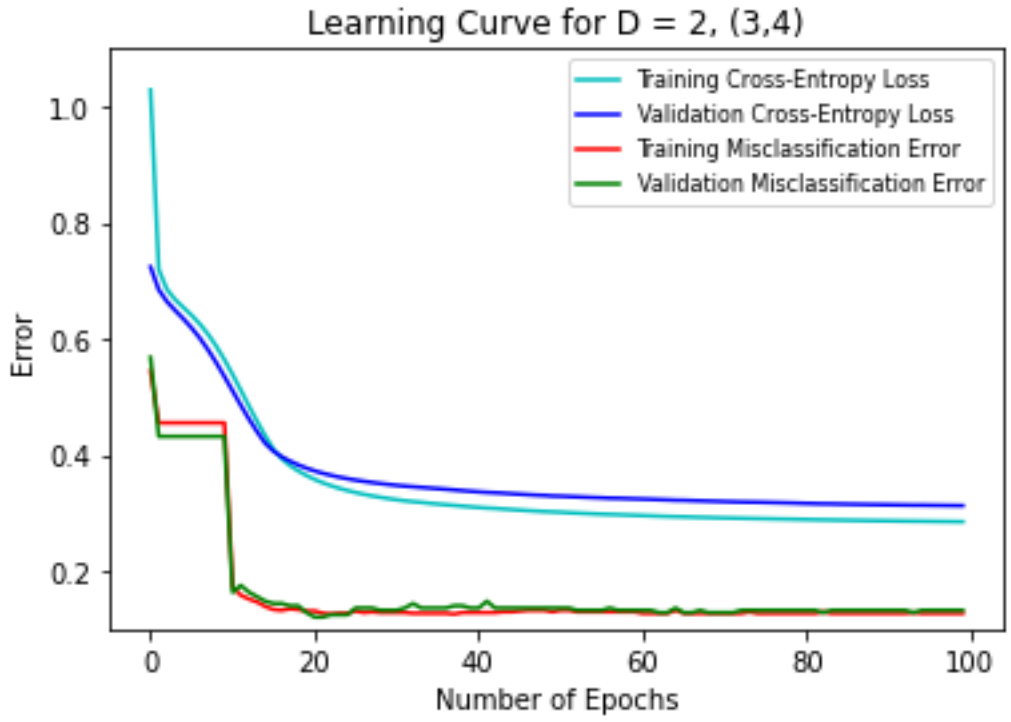


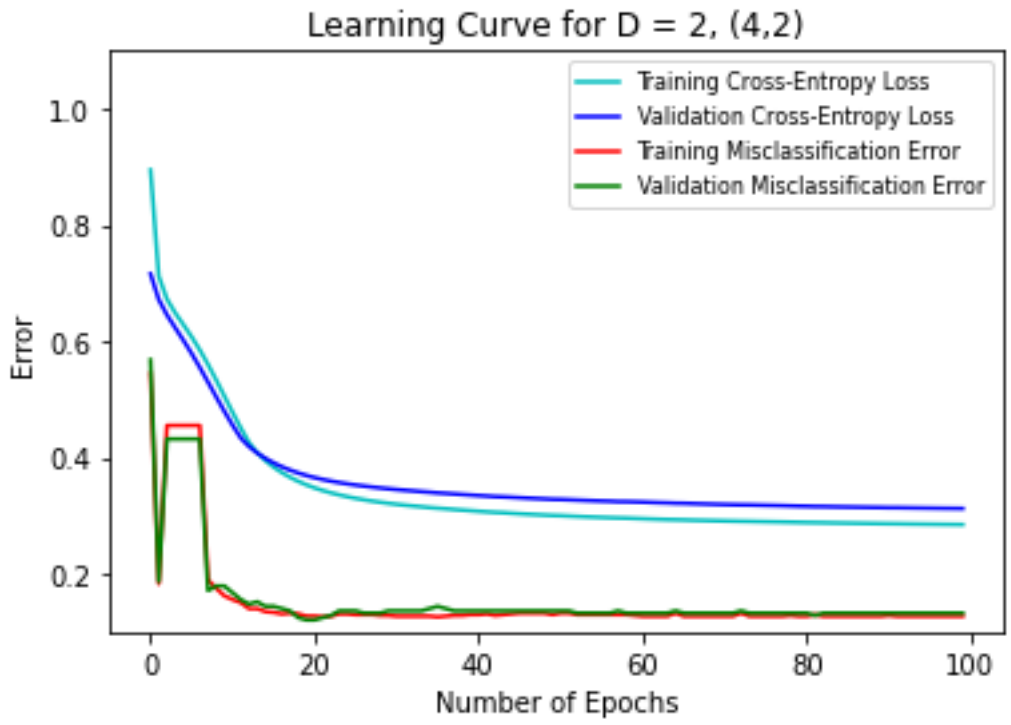


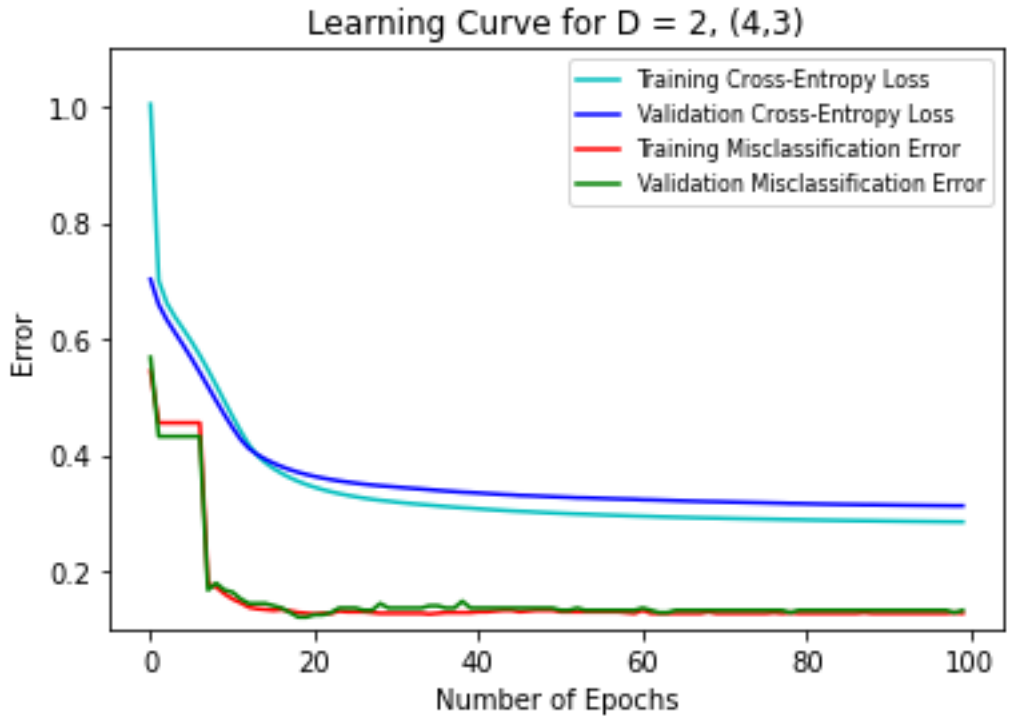

Learning Curve for D = 2, (2,4)
Training Cross-Entropy Loss
Validation Cross-Entropy Loss
Training Misclassification Error
Validation Misclassification Error
Error
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
Number of Epochs

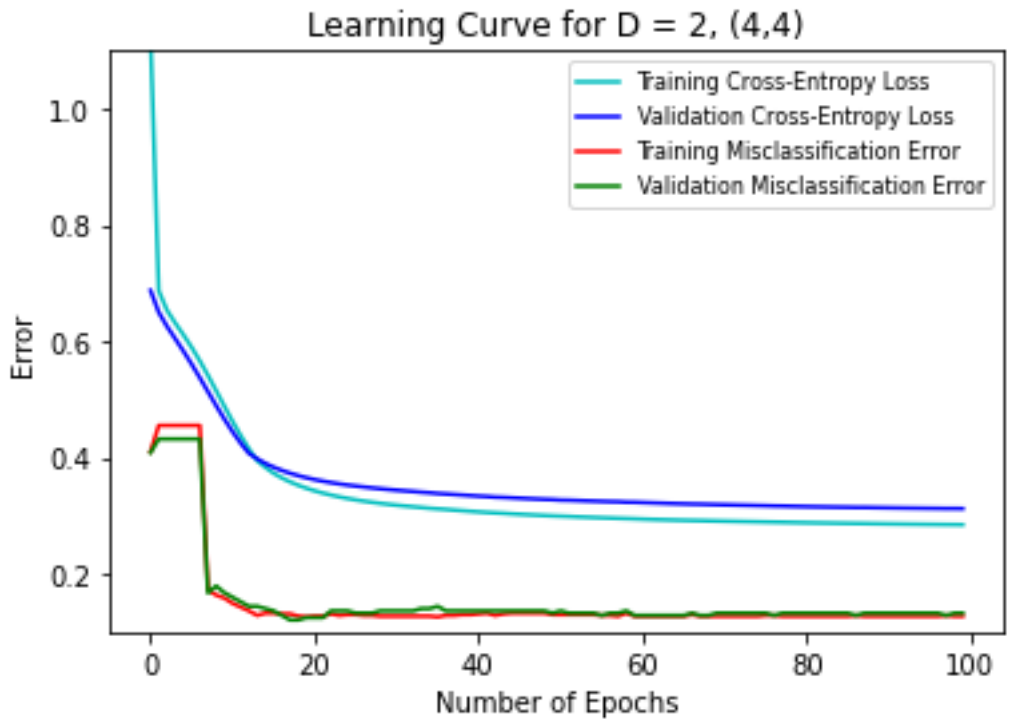

Learning Curve for D = 2, (3,2)
Training Cross-Entropy Loss
Validation Cross-Entropy Loss
Training Misclassification Error
Validation Misclassification Error
Error
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
Number of Epochs


Learning Curve for D = 2, (3,3)
Training Cross-Entropy Loss
Validation Cross-Entropy Loss
Training Misclassification Error
Validation Misclassification Error
Error
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
Number of Epochs


Learning Curve for D = 2, (3,4)
Training Cross-Entropy Loss
Validation Cross-Entropy Loss
Training Misclassification Error
Validation Misclassification Error
Error
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
Number of Epochs


Learning Curve for D = 2, (4,2)
Training Cross-Entropy Loss
Validation Cross-Entropy Loss
Training Misclassification Error
Validation Misclassification Error
Error
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
Number of Epochs


Learning Curve for D = 2, (4,3)
Training Cross-Entropy Loss
Validation Cross-Entropy Loss
Training Misclassification Error
Validation Misclassification Error
Error
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
Number of Epochs



Discussion

The overall trend in all figures show that both training and validation cross-entropy losses and the training and validation misclassification errors have a gradual decrease in the first few epochs and then slowly stabilize to a value as number of epochs increase. The curves are nonincreasing as after every update, the model accuracy becomes better. There is a bit of some fluctuations in the errors however that is due to noisy performance throughout the training.

To analyze deeper into the curves, the effect of the number of units in the hidden layers can be analyzed. It is seen that as the number of units in hidden layer 2, n2, the starting training cross-entropy loss is greater with little to no change in the steepness of the curves. However, the shapes stay fairly the same. When the number of units in hidden layer 1, n1 increases there is a more noticeable change between the plots. When n1 increase, the losses have a steeper decrease, i.e., decreasing faster in a smaller number of epochs. The pattern applies to the errors as well – the errors decreases faster with a smaller number of epochs.

The training cross-entropy loss starts at a higher value than the validation cross-entropy loss in the first few epochs and eventually goes below the validation cross-entropy loss. The same pattern applies for the misclassification errors. It starts with a higher training error than the validation error and then eventually goes below the validation error. This is expected because that means the point of intersection between the errors mean that overfitting is likely starting to occur. The validation set is brought into the training of the model because it helps with identifying when model starts to overfit, so that training can be stopped. Without the validation set, it is possible for the training error to reach zero. However, this can affect the prediction on unseen data. By knowing when the model starts to overfit, this helps with the overall accuracy of the model.